Computer Science and Application 计算机科学与应用, 2023, 13(3), 458-464
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133044

Hans汉斯

多模态行人重识别系统研究

常　文，游福成

北京印刷学院信息工程学院，北京

收稿日期：2023年2月18日；录用日期：2023年3月20日；发布日期：2023年3月28日

摘　要

行人重识别系统用来找寻行人，有着举足轻重的作用，而且还可以利用语义快速从库中查找出最相似的人。这种应用基于图像和文本的不同模型结构，配上合适的损失函数，将模型收敛。现基于多模态学习，可以使用语义进行图像的搜寻，对于数据量极大的监控系统而言，这无疑能帮助更高效的找寻目标。在文本分类的模型中，深度学习模型常出现在人们视野中，但是由于模型深度等原因，深度学习框架往往时间复杂度比较高，而FastText模型是基于嵌入的模型，没有复杂和深度的框架，但是却能在保证准确性的同时大幅提高模型训练的速率，使行人查找搜寻任务可以更快被完成，推进相关行业发展。因此，本文将FastText模型应用于多模态行人重拾别系统研究时，显著提高了训练的速度，将多模态行人重识别系统推向了更好的应用层面。

关键词

行人重识别，深度学习，FastText

Research on Multi-Model Pedestrian Re-Identification System

Wen Chang, Fucheng You

College of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Feb. 18th, 2023; accepted: Mar. 20th, 2023; published: Mar. 28th, 2023

Abstract

The pedestrian re-identification system is used to find pedestrians, which plays a pivotal role, and can also use semantics to find the most similar person from the dataset quickly. This application is based on different model structures of images and text, with appropriate loss functions, to converge the model. Now based on multi-model learning, you can use semantics for image search, which can undoubtedly help to find targets more efficiently for monitoring systems with a large

文章引用: 常文, 游福成. 多模态行人重识别系统研究[J]. 计算机科学与应用, 2023, 13(3): 458-464.
DOI: 10.12677/csa.2023.133044

amount of data. In the model of text classification, deep learning models often appear in people's field of vision, but due to model depth and other reasons, deep learning frameworks tend to have high time complexity, while model of FastText is based on embedding model, without complex and deep frameworks, but can ensure accuracy while improving the speed of model training, so that pedestrian re-identification tasks can be completed faster, promoting the development of related industries. Therefore, when the FastText model is applied to the research of multi-model pedestrian re-identification system, the training speed is significantly improved, and the multi-model pedestrian re-identification system is pushed to a better application level.

Keywords

Pedestrian Re-Identification, Deep Learning, FastText

1. 引言

现如今，城市里的很多公共场所都配有监控摄像头，甚至每秒钟都能产生成千上万的视频数据，利用这些数量庞大的数据去搜寻可能的犯罪嫌疑人，现阶段这项工作还是利用人工居多，并且可能需要消耗数十天甚至数月的时间来完成搜寻工作。因此，自动搜寻行人的方法是我们非常迫切需要的。使用不限格式的文本描述在庞大的数据集中搜寻行人图片在计算机视觉领域是一个非常具有挑战性的领域。我们利用表征模型将图像和文本投影到相同的表征空间之中，再利用损失函数计算它们的相似性，得出损失，利用反向传播收敛模型。经过训练，得出的最优模型。模型最终目的是为了输入不限格式的文本描述，在图像数据集中搜寻相关性最高的行人图片。

图像和文本在自身的模态内都包含有丰富的语义信息，为了能够用同一个模型来挖掘对比信息，匹配图像－文本对是我们面临的挑战。基于文本的行人搜寻目的是根据所给出的文本描述的句子来搜寻目标人物的图片。这个内容的要求就是将两个模态的内容使用强表征联系起来。强鲁棒性的表征是将图像和文本投影到某同一个向量空间而形成，再利用特定的损失函数计算出损失进行反向传播来训练模型。图像的特征提取网络包含有：CNN、VGG16、Yolo3、ResNet、ViT 等网络模型，文本的特征提取包括：CNN、DPCNN、VDCNN、RNN、LSTM、BiLSTM、GRU、BiGRU、Bert 等网络模型，现阶段大多数使用的损失函数包括有：CMPM、RankingLoss 等等。同时为了使模态内表征更强，在图像方面特征增强的方法包括：CAM、PAM、Max Gated block、CIA Module、SIA Module 等等，在文本数据增强方面的方法包括：EDA、回译等方法，在图像数据增强方面，包含一些基础的数据增强的方法包括：随机裁剪、随机翻转等等方法，在文本嵌入方面的方法包括：Word2Vec、GloVe、Bert 等嵌入方法。

由于在找寻行人的时候，现实情况中缺少行人的图像，所以用行人图像来找人的方式不切实际；而基于自然语言描述，在行人图集中寻找最匹配的行人，是一种比较贴合实际的方式；在应用方面：找犯罪嫌疑人、视频监控或者寻人启事等事务之中，此研究内容具有很强的应用性和价值。

2. 相关工作

在使用文本描述的行人重识别方法方面，研究者们提出了许多方法。其中根据学习方式可以划分为：模态间以及模态内两种方法；其中根据图像特征划分可以分为：局部分支匹配、全局分支匹配以及混合

匹配；根据多模态特征的特征维度可以划分为：高纬特征匹配、低维特征匹配以及混合匹配；根据损失函数也可以划分为：基于标签的以及基于实例的；还有根据文本嵌入方式可以划分为：嵌入为单词形式的以及嵌入为字母形式的，等等。

这些划分方式，都起到了不同的效果。例如通过将图像表征进行划分，例如使用：PCB [1]的图像表征划分方式，再与文本的表征进行匹配，强调突出了表征信息突出的部分，而全局的分支匹配则保留了如背景等容易被忽略的图像表征信息。Sun [1]等人水平分割行人图像的特征矩阵来挖掘局部表征，突出特征将被更好的学习；Wang [2]等人采用多尺寸表征学习策略：将每个特征矩阵剪切成尺寸大小不同的矩阵作为局部分支进行学习；Zhao [3]等人使用了特殊的注意力机制进行学习，进一步增强图像局部表征的鲁棒性；Song [4]等人以及 Zhao [5]等人使用了图像姿态评估特征辅助局部分割特征以及挖掘细节信息；Kakayeh [6]等人使用附加图像语义分割的方法来增强图像的局部特征；Zhang [7]等人 CAM 和 PAM 模块加入到 ResNet50 中间层来增强图像表征；Ma [8]等人使用 Max Gated block 来减轻 GMP 中引入的错误信息来提高行人重识别模型的准确率；Hou [9]等人使用 CIA 和 SIA 模块加入到 ResNet50 中来增强图像的表征；Li [10]等人提出 GNA-RNN 模型，提取多模态特征进行拼接，使用了多级别 attention 机制等特征混淆技术；Zhang [11]等人使用 LSTM-CNN 模型提取特征，提出了 CMPC 等损失函数进行训练。

我们的基准模型，使建立再 CMPC 损失函数和图像使用 ResNet 残差网络，语义使用 BERT 预处理后使用残差分支来将图像和语义数据的表征化。损失函数使用 CMPC。而我们的模型，将语义的网络修改成 FastText 模型，牺牲了部分精准度，但是完成了时间复杂度上的大幅减少，是一种非常容易落地和解决现实应用问题的具体方案。

3. 基准模型

这一节中，我们会介绍一下基准模型：基于部分的文字图像卷积网络。此模型阐述了一个双向的对齐网络框架，基于图像 CNN 分支和文本 CNN 分支，并且使用 CMPC 损失函数来收敛模型。

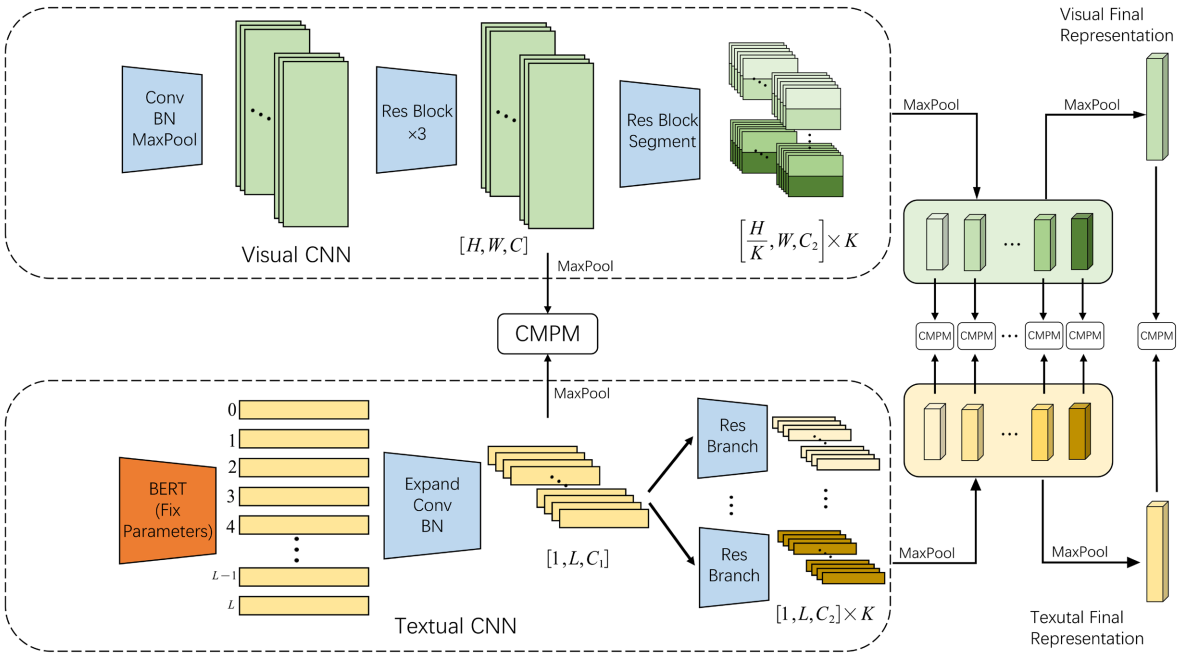

Figure 1. Graphic of the basic model
图 1. 基准模型图

3.1. 图像表征学习

如图 1 所示，模型包含两个 CNN 分支来学习图像和语义(输入的行人图像和描述)表征之间的差别，在训练姐阶段，假设训练集为 $D=\{I_i,T_i\}_{i=1}^{N}$，其中 N 表示每一批的图像文本对，每一对是由图像 I 和对应的描述 T 组成。在图像 CNN 分支中，采用 ResNet-50 框架作为挖掘图像特征的基础框架，由四个残差块组成。不同的残差块可以从不同的维度来捕获不同的语义信息。对于每个图像 I，我们定义由第三个和第四个残差块作为低纬特征图 $f_l^I\in R^{H\times W\times C_1}$ 以及高维特征图 $f_h^I\in R^{H\times W\times C_2}$，其中 H，W 以及 C_1/C_2 代表特征图的高度和宽度的维度以及通道数，因此，我们可以使用全局最大池化层作为过滤器来挖掘显著信息。如下公式：

$$v_l^I=GMP\left(f_l^I\right) \tag{1}$$

同时，采用了 PCB [2]策略获取本地分支。特别需要注意的是，高纬特征图 f_h^I 被分割为 K 个水平块 $\left\{f_{p1}^I,f_{p2}^I,\cdots,f_{pK}^I\right\}$，其中 f_{pi}^I 的维度为 $(H/K)\times W\times C_2$。对于每个块也使用最大池化层挖掘图像本地表征 $v_{pi}^I\in R^{C_2}$。为了混淆所有本地表征，使用通道维度元素中最大的值，以此获得图像全局表征 $v_g^I\in R^{C_2}$：

$$v_g^I=Max\left(v_{p1}^I,v_{p2}^I,\cdots,v_{pK}^I\right) \tag{2}$$

因此，我们可以获得图像特征集 $V^I=\left\{v_l^I,v_{p1}^I,\cdots,v_{pK}^I,v_g^I\right\}$，其中包含了低维、本地和全局表征。在测试阶段，则只使用高位表征来测试相似性。

3.2. 文本表征学习

在文本分支，BERT [12]是一种表现特别好的模型，我们将其用在挖掘特殊字嵌入上，通过双向 Transformer [13]训练，可以学习到文本上下文之间的关系。特别是，将文本描述 T 分割成字的集合，然后插入[CLS]和[SEP]在集合开头和结尾。然后，这个集合使用预训练分词器将集合嵌入到表征中。为了确定文本长度的一致性，当文本长于 L 的时候，选择前 L 个表征，当文本短于 L 的时候尾部加上 0 补齐。最后，每个分词后的文彪表述被输入进预训练以及修成参数后的 BERT 模型之中，来挖掘字嵌入 $t\in R^{L\times D}$。其中 D 代表每个词嵌入的维度。由于与训练后的 BERT 模型有强语义表征能力，并且文本 CNN 结构有足够的能力来处理字嵌入，同时只有 CNN 结构才能显著的减少参数，加速模型聚拢。

为了能满足卷积层的输入要求，将字嵌入从 $t\in R^{L\times D}$ 扩展到 $t^*\in R^{1\times L\times D}$，其中 1，$L$ 和 D 分别代表高度、宽度和卷积输入的通道数。由于残差网络和文本 CNN 等结构，设计了一种多分枝文本 CNN 如图 2。在文本 CNN 中，为了能将字嵌入散列成预图像低纬特征图相同通道维度，第一个卷积层的过滤器被设置为 $1\times 1\times D\times C_1$。然后我们可以得到文本地位特征图 $f_l^T\in R^{1\times L\times C_1}$。

多分支文本 CNN 包括 K 个残差分支，分别对应图像中行人照片的 K 块。对于每个分支，包含了 P 个文本残差模块，目的是自适应学习文本表征来匹配图像本地表征。文本模块与 ResNet 中有同样的模型，由几个卷积层和批量正则层组成。短接被应用于从低纬层传递信息给高维层，扭转了网络梯度下降问题并加速了模型训练。特别是，保持文本信息不被压缩，所有残差模块中的卷积层的步长被设置为 1 × 1。对第一个残差模块的每一个分支，修改文本特征图的通道维度为 C_2，与图像高维特征图 $f_h\in R^{H\times W\times C_2}$，并且保持通道维度不变进入下一个残差模块。经过多分支文本 CNN 后，可以获得文本本地特征图。同样的，采用同图像 CNN 分支一样的最大池化层挖掘文本本地特征并选择每个向量维度元素的最大值来混淆这些本地表征。然后，可以得到 $V^T=\left\{v_l^T,v_{p1}^T,\cdots,v_{pK}^T,v_g^T\right\}$，包含低维、本地和全局表征。

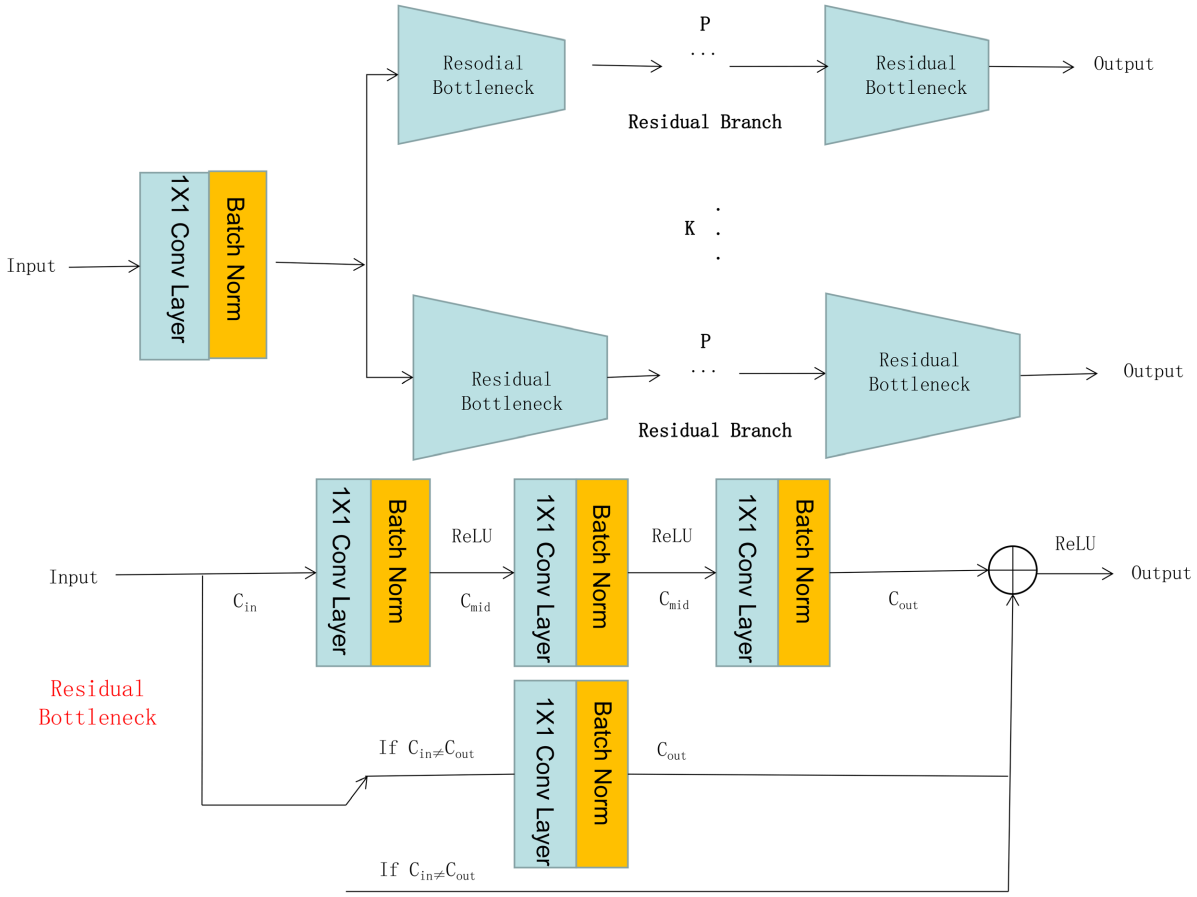

Figure 2. Graphics of residual branch and residual bottleneck
图 2. 残差分支及残差瓶颈图

3.3. CMPC 损失函数

根据图像和文本的表征集合 V^I 和 V^T，有损失函数如下：

$$Loss = \lambda_1 L^l_{CMPM} + \lambda_2 \sum_{k=1}^{K} L^{pk}_{CMPM} + \lambda_3 L^g_{CMPM} \tag{3}$$

其中，$\lambda_1, \lambda_2, \lambda_3$ 表示控制不同 CMPM 损失的重要程度，并且 $\lambda_1 L^l_{CMPM}$，$\sum_{k=1}^{K} L^{pk}_{CMPM}$，$L^g_{CMPM}$ 分别表示 CMPM 损失中低维、本地以及高维表征的损失。具体损失函数内容，可以看 Zhang [11]的文章。

4. FastText 模型

FastText 是一种简单而且高效的文本分类和字嵌入方法，分类的精度在某些方面甚至能够媲美深度学习，而由于模型复杂度不高，因此在训练和测试速度上有很大的优势。其核心思想是 N-gram，一种基于统计语言模型的算法，也是 NLP 领域的一种较为特殊的嵌入模型。在文本特征提取中，基本思想是将文本内容按照顺序进行窗口大小为 N 的滑动操作，最终形成长度为 N 的片序列。这种滑动窗口的模式包含字粒度和词粒度两种形式。而使用 FastText 进行文本分类的同时会产生词的嵌入，我们正是想利用这种高效且有效的方式来进行多模态行人重识别系统的研究。FastText 模型如下图 3，其中 x_i 代表 i-Gram 嵌入：

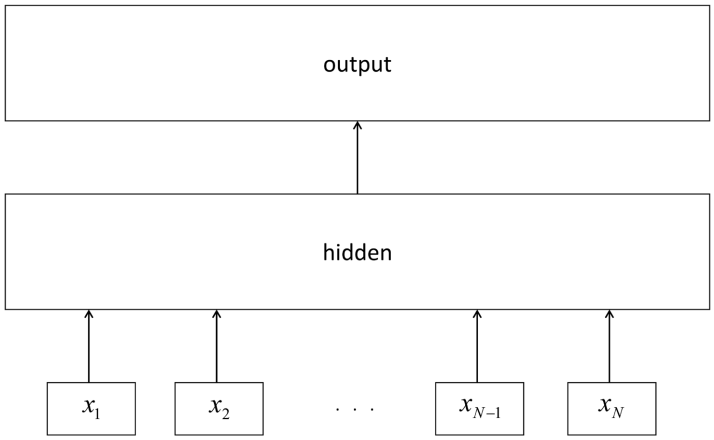

Figure 3. Structure of FastText
图 3. FastText 结构图

基于 FastText 速度快、精度高的特性，使用 FastText 代替文本 CNN 进行词嵌入，利用 FastText 模型进行 N-Gram 叠加并于词嵌入做并使用全链接层，作为多模态学习的文本表征，按照图像的低维、本地以及全局的三种形式进行全链接，保障于图像的三类表征维度相同，实现文本和图像表征的对齐，以此来训练。在使用 CMPC 损失函数进行反向传播，使模型收敛。

5. 实验

5.1. 实验设置

数据集使用 CUHK-PEDES 数据集[10]，只包含对人的描述和行人图像。包含 40,206 个行人图像和 13,003 个实体。对于每张行人照片，对应有两句描述，每句话描述大约 23 个字，词汇表有 9408 个不同的字。

评价方式采用标准评价衡量指标，使用 top-k (k = 1, 5, 10)准确度以及每轮训练时间来评价模型表现。

实现细节，在图像 CNN 分支中，采用预训练后的 ResNet-50 作为基础挖掘图像特征图，步长做微调整；在 FastText 文本分支，使用本身与 2-Gram 和 3-Gram 进行拼接和取均值，再使用线性层，使用 BERT 预训练框架进行字嵌入。所有的输入图像被重置成 384*128，规定文本长度 $L=64$。本地分隔值 $K=6$。在图像和文本特征里，H 被设置为 24，W 被设置为 8，$D=768$，$C_1=1024$，$C_2=2048$。批数量设置为 64 对图像文本对。

训练阶段，把 Adam 作为优化器，权重损失值设置为 3×10^{-3}，模型训练 80 个循环。基础学习率为 3×10^{-3}，50 轮后减少 0.1。在前十轮的学习中，使用 warmup 策略来初始化学习率。三个超参数 $\lambda_1,\lambda_2,\lambda_3$ 都设置为 1。实在使用 pytorch 框架，GPU 使用 tesla v100 (16G)。

5.2. 实验结果对比

Table 1. System resulting data of standard experiment
表 1. 标准试验系统结果数据

模型	一轮学习时间	Top1	Top5	Top10
Visual CNN + Text CNN	12 min	0.636	0.828	0.891
Visual CNN + FastText	**4 min**	0.526	0.734	0.796

对照组和实验组，两组实验都进行了 80 轮学习，有如表 1 数据。使用 Visual CNN 和 Text CNN 进行特征挖掘的模型，其一轮学习时间大约为 12 min，top1 准确度 63.6%，top5 准确度 82.8%，top10 准确度 89%；而我们提出的模型，使用 Visual CNN 和 FastText 进行特征提取，在一轮学习时间上只花了 4 min，速度提升了快两倍，也保证了一定的准确度，top1 准确度 52.6%，top5 准确度 73.4%，top10 准确度 79.6%。

6. 结论

通过对比原 Visual CNN 和 TextCNN 的行人重识别系统，使用 Visual CNN 和 FastText 的模型虽在 top1、top5 和 top10 的精度上下降了 15%左右，但是时间复杂度上，后者将速度提上了一个平台，提升了 200%左右(学习时间为保留分钟位)。因此在平常的应用中，本研究证明了可以利用 Visual CNN 和 FastText 进行图像和文本的特征挖掘，并对齐后使用 CPMC 损失函数进行学习，对比于对照组，大幅提升了训练和测试的速度，降低了参数量，并且保证了一定的精度。此研究在一定程度上会促进项目的落地和应用。

参考文献

[1] Sun, Y., Zheng, L., Yang, Y., Tian, Q. and Wang, S. (2018) Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). *The European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 480-496.

[2] Wang, G., Yuan, Y., Chen, X., Li, J. and Zhou, X. (2018) Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *Proceedings of the* 26*th ACM International Conference on Multimedia*, Seoul, 22-26 October 2018, 274-282.

[3] Zhao, L., Li, X., Zhuang, Y. and Wang, J. (2017) Deeply-Learned Part-Aligned Representations for Person Re-Identification. *Proceedings of the IEEE International Conference on Computer Vision*, *Venice*, 22-29 October 2017, 3219-3228.

[4] Song, G., Leng, B., Liu, Y., Hetang, C. and Cai, S. (2018) Region-Based Quality Estimation Network for Large-Scale Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 32.

[5] Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X. and Tang, X. (2017) Spindle Net: Person Re-Identification with Human Body Region Guided Feature Decomposition and Fusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 1077-1085.

[6] Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E. and Shah, M. (2018) Human Semantic Parsing for Person Re-Identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 1062-1071.

[7] Zhang, S., Yin, Z., Wu, X., Wang, K., Zhou, Q. and Kang, B. (2021) FPB: Feature Pyramid Branch for Person Re-Identification. arXiv preprint arXiv: 2108.01901.

[8] Ma, T., Yang, M., Rong, H., Qian, Y., Tian, Y. and Nabhan, N. (2021) Dual-Path CNN with Max Gated Block for Text-Based Person Re-Identification. *Image and Vision Computing*, **111**, 104168.

[9] Hou, R., Ma, B., Chang, H., Gu, X., Shan, S. and Chen, X. (2019) Interaction-and-Aggregation Network for Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-20 June 2019, 9317-9326.

[10] Li, S., Xiao, T., Li, H., Zhou, B., Yue, D. and Wang, X. (2017) Person Search with Natural Language Description. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1970-1979. https://doi.org/10.1109/CVPR.2017.551

[11] Zhang, Y. and Lu, H. (2018) Deep Cross-Modal Projection Learning for Image-Text Matching. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 686-701.

[12] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv: 1810.04805.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *The International Conference on Neural Information Processing Systems* (*NeurIPS*), Long Beach, 4-7 December 2017, 6000-6010.